Knowledge-Based Systems 295 (2024) 111683

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Supervised Cross-Momentum Contrast: Aligning representations with prototypical examples to enhance financial sentiment analysis

Bo Peng [*], Emmanuele Chersoni, Yu-yin Hsu, Le Qiu, Chu-ren Huang

*Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*

ARTICLE INFO

Dataset link: https://github.com/PengBO-O/SuCroMoCo

*Keywords:*
Financial sentiment analysis
Pre-trained language models
Supervised contrastive learning
Cross-momentum contrast

ABSTRACT

Financial sentiment analysis plays a pivotal role in understanding market dynamics and investor sentiment. In this paper, we propose the **Su**pervised **Cro**ss-**Mo**mentum **Co**ntrast (SuCroMoCo) framework, a novel approach for financial sentiment analysis. SuCroMoCo leverages supervised contrastive learning and cross-momentum contrast to align financial text representations with prototypical representations based on sentiment categories. This alignment greatly improves classification performance, addressing the limitations of pre-trained language models (PLMs) in fully grasping the intricate nature of financial text. Through extensive experiments, we demonstrate that SuCroMoCo outperforms existing PLMs-based approaches and Large Language Models (LLMs) on diverse benchmark datasets.

## 1. Introduction

Financial Sentiment Analysis (FSA) is a growing research field that focuses on extracting emotion polarities from financial language texts [1]. FSA plays a critical role in assisting investors in making well-informed investment decisions by analyzing a wide range of information sources, such as financial news, social media sentiment, and other data to identify market trends [2,3], detect fraud [4,5], and identify emerging risks [6–8], among other applications. The advent of pre-trained language models (PLMs) [9–12] has significantly advanced various natural language understanding tasks, including sentiment analysis [13,14]. As a result, there is a growing interest in harnessing the potential of PLMs to enhance the efficiency of FSA in the financial domain. A widely adopted approach involves using PLMs as encoders to produce generic text representations, followed by fine-tuning a classifier on top of these representations for classification. This method has demonstrated successful performance on numerous benchmark datasets [15]. However, it is important to acknowledge that PLMs may lack specialized knowledge in financial language texts. Consequently, they may produce suboptimal representations, which can have a detrimental impact on their performance in FSA tasks. Financial texts, such as corporate reports and earnings calls, are often lengthy and filled with technical terminology that requires expertise in the specific domain [16,17].

A practical approach is adapting a language model to the financial domain using self-supervised pre-training [9]. This adaptation process involves training the models either from scratch or by continuously pre-training a language model that was originally trained on a general domain corpus [17–19]. By leveraging self-supervised pre-training, language models can acquire valuable semantic information specific to financial texts and encode it into their representations.

However, self-supervised pre-training, while enhancing semantic understanding, may not necessarily lead to improved accuracy in financial sentiment classification. Peng et al. [20] found that domain-specific PLMs may still exhibit lower performance compared to general domain PLMs in financial sentiment analysis tasks, even after supervised fine-tuning. This discrepancy can be attributed to the unique set of sentiment expressions in financial texts, which extend beyond emotions and opinions and are closely intertwined with investors' expectations and perceptions of favorable or unfavorable directions for stocks, events, or financial instruments [21–23]. The intricate relationship between sentiment and potential financial outcomes implies that certain words and phrases may have varying sentiment polarities in a financial context. For example, originally neutral financial concepts may acquire negative or positive connotations when combined with verbs and directional information. The models might not fully capture these nuances of sentiment variations from the learned contextual representations of financial texts, leading to suboptimal performance. Consequently, additional processing is required to effectively correlate financial context with sentiment polarity.

To overcome the challenge of adapting PLMs to the nuanced sentiment expressions in financial texts, our proposed solution involves

[*] Corresponding author.
*E-mail addresses:* bopeng@polyu.edu.hk (B. Peng), emmanuele.chersoni@polyu.edu.hk (E. Chersoni), yu-yin.hsu@polyu.edu.hk (Y. Hsu), lani.qiu@connect.polyu.hk (L. Qiu), churen.huang@polyu.edu.hk (C. Huang).

https://doi.org/10.1016/j.knosys.2024.111683
Received 9 October 2023; Received in revised form 22 February 2024; Accepted 19 March 2024
Available online 22 April 2024

aligning financial text representations with *prototypical examples of sentiment polarity classes*. These prototypical examples are chosen to share identical sentiment polarities with the target texts, featuring a clearer and unequivocal sentiment expression that enables accurate classification post fine-tuning. In essence, our approach strives to bring financial representations closer to those that are prototypical for the same sentiment category, while simultaneously pushing them farther away from prototypes with a different sentiment polarity. This alignment process serves to reshape the representation space, fostering proximity among examples with the same sentiment polarity and creating distinct boundaries with those of different polarities.

To elucidate further, we aim to enhance the model's effectiveness in encoding financial texts without the resource-intensive process of extensive domain knowledge learning. Instead, we introduce prototype data and leverage contrastive learning techniques. This strategic combination not only facilitates the efficient encoding of financial texts but also contributes to the discriminative classification of financial examples. For our experiments, we selected the prototypical examples from datasets with a high level of inter-annotator agreement, to ensure that the polarity of such examples is clear and unambiguous for humans. These datasets can either be from the same financial domain or from a general domain; in the latter scenario, our method can be seen as a form of domain adaptation.

To achieve our goals, we introduce **Su**pervised **Cro**ss **Mo**mentum **Co**ntrast (**SuCroMoCo**), a simple yet efficient framework, to achieve the desired alignment. Momentum contrast (**MoCo**) [24] is employed as the underlying contrastive learning method, which has been proven successful in unsupervised representation learning tasks. We generalize it to supervised learning [25,26] by maintaining the label queue that corresponds with the representation queue. Unlike the standard contrastive learning setting that commonly uses self-augmentation examples for contrastive learning [27,28], we consider financial text representations and prototype representations as augmentations for each other. We employ cross-contrast between financial text representations and prototype representations. With the representation queue and momentum update, a single financial or prototypical text representation is contrasted with a rich set of prototypical or financial text representations simultaneously. This approach allows the financial text representations to move closer to the overall distribution of prototype representations belonging to the same sentiment category, rather than being limited to a specific representation. As shown in Fig. 1, the upper figure indicates that the prototype representations can be accurately classified after fine-tuning, while the financial text representations still cannot be accurately classified. In contrast, our proposed method, SuCroMoCo, leverages supervision information across financial and prototypical texts, gathering financial text representations closer to prototype representations, and resulting in more efficient classification. It is important to note that this paper specifically focuses on applying SuCroMoCo to the task of financial sentiment analysis; nonetheless, this can be applied to other classification problems that involve the availability of prototypical examples for the target classes.

In order to validate the effectiveness of SuCroMoCo, we conducted comparison experiments on three benchmark datasets for Financial Sentiment Analysis (FSA). In our experiments, we trained SuCroMoCo using the pre-trained BERT and RoBERTa as foundational models. To ensure a comprehensive evaluation, we compared SuCroMoCo with various baselines, including several PLMs pre-trained on financial text, and four recently-introduced Large Language Models with in-context learning abilities. Remarkably, SuCroMoCo consistently outperformed all the other methods in sentiment analysis for financial texts, highlighting its superiority and effectiveness in this domain. To gain a deeper understanding of how SuCroMoCo achieves such high performance, we provide visualizations of the learned representations and conducted in-depth ablation studies. Our analyses provide additional insights into the alignment process and about the individual contributions of each component in SuCroMoCo, shedding light on its inner workings and advantages.

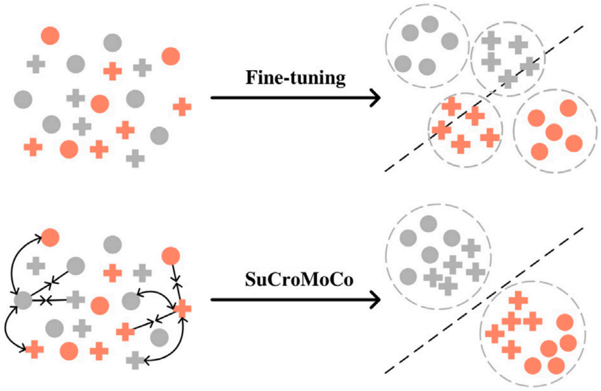


**Fig. 1.** The concept of our proposed SuCroMoCo framework. Each color represents one class. The circle and plus denote the representation of the prototypical example and financial text respectively.

## 2. Related work

### 2.1. Financial sentiment analysis

In financial contexts, specific words can undergo polarity shifts compared to their general language usage. For instance, Loughran and McDonald [29] found that 73.8% of commonly used negative words adopt a neutral stance when used in financial contexts. Chen et al. [30] also observed this phenomenon in social media texts. Furthermore, Xing et al. [31] introduced a cognitive-inspired sentiment lexicon, employing an exploration–exploitation mechanism to balance the discovery of new sentiment words with the updating of polarity scores.

During the early stages of financial sentiment analysis, lexicon-based approaches were extensively employed. These methods determine sentiment polarity by aggregating individual words or phrases from sentiment lexicons. Moreover, lexicon-based methods often incorporate considerations of word weighting and utilize statistical features like count vectorizers and TF-IDF to learn feature representations. These approaches can be synergized with traditional machine learning techniques, such as SVM and Logistic Regression, to enhance performance [32,33]. However, one limitation of these methods is their potential oversight of contextual features in sentences, which play a pivotal role in capturing the authentic sentiment expressed in texts.

In contrast, deep neural networks such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks [34] were regarded as better choices for capturing both semantic and contextual information. Such neural networks function as encoders, learning sentence representations in end-to-end architectures [35], and they were often applied to the problem of analyzing sentiment in financial texts [36,37].

More recently, Transformer-based pre-trained language models [38] have demonstrated remarkable capabilities in understanding natural language texts. After being pre-trained on extensive corpora through self-supervised tasks (e.g. masked language modeling, next sentence prediction etc.) PLMs serve as robust encoders, generating highly informative sentence representations and significantly improving the performance of sentiment analysis [15,39].

Additionally, these self-supervised pre-training tasks can be applied to adapt PLMs to financial domain and yield performance gains in FSA. This adaptation can be achieved by employing a second-phase pre-training, known as in-domain (domain-adaptive) pre-training [40,41], or by training language models from scratch on domain-specific corpora. Araci et al. [18] observed that performing in-domain pre-training exclusively on target datasets could lead to improvements. Yang et al. [17] adopted the same self-supervised pre-training tasks on a large scale of financial communication corpora to train financial domain-specific BERT (FinBERT), either from scratch or through continuous pre-training based on BERT. Shah et al. [19] applied a strategy of masking financial keywords and phrases, along with span boundary and in-filing objectives, to pre-train financial language models.

However, upon evaluating the effectiveness of the aforementioned financial domain-specific PLMs across various tasks, including FSA, Peng et al. [20] found that the improvements achieved were not consistently significant.[1] These findings are also echoed in the experiments conducted in the present study (as detailed in Section Section 5.1). Such observations underscore the notion that self-supervised in-domain adaptation of PLMs to the financial domain, even with an improved understanding of financial context, might not directly lead to substantial gains in sentiment classification accuracy for financial texts. Consequently, in this study, we propose a more straightforward approach. We intend to utilize contrastive learning techniques to align the representations of financial text and prototypical examples within the embedding space. This alignment will be guided by categorical information, potentially offering a more robust means of enhancing sentiment analysis for financial texts.

### 2.2. Contrastive learning

Contrastive learning has emerged as a competitive technique for representation learning, employing self-supervised losses rooted in noise-contrastive estimation (NCE) [43]. This approach encourages similar sample pairs (positive pairs) to cluster together while pushing dissimilar sample pairs (negative pairs) farther apart within the learned representation space. To this end, Oord et al. [44] introduced the Information Max Noise Contrastive Estimation (InfoNCE) loss. This loss maximizes the model's ability to capture information that predicts neighboring examples effectively, improving model's ability to learn valuable representations of visual and textual data. Expanding upon InfoNCE, Chen et al. [27] proposed the normalized temperature-scaled cross-entropy loss (NT-Xent), a core element of the SimCLR framework designed for visual representation learning. SimCLR accentuates semantically related images and their augmentations, often achieved through augmentation techniques like cropping, rotation, and cutout.

Such strategies have also been adapted to textual data augmentation. Yan et al. [45] explored various augmentation techniques, such as adversarial attacks, token shuffling, cutoff, and dropout, to create contrasting views of sentences. Gao et al. [28] introduced SimCSE, which involves sentence augmentation by inputting the same sentence twice into the PLM encoder. Chuang et al. [46] employ stochastic masking and sampling from a masked language model to create augmentation pairs.

Beyond augmentation strategies, the choice of a large batch size during contrastive training significantly impacts performance [27]. However, managing such large batches can be resource-intensive and even computationally impractical for encoder updates via back-propagation. To alleviate it, He et al. [24] proposed momentum contrast (MoCo), which treats contrastive learning as a dynamic dictionary look-up. In MoCo, a momentum-updated encoder is introduced to maintain a large size queue of negative examples for contrastive learning. Building upon this, Wu et al. [47] enhanced SimCSE by integrating it with MoCo, enlarging the number of negative pairs to further refine the contrastive learning process.

Contrastive learning extends naturally to the fully supervised setting to leverage label information more effectively. Differing from the self-supervised scenario where only augmented input examples are contrasted, supervised contrastive learning pairs all examples from the same category as positives against a distinct set of negative examples [25]. Gunel et al. [26] combined supervised contrastive loss with cross-entropy loss to fine-tune PLMs, resulting in performance enhancements for text classification tasks. Dai et al. [48] transformed supervised contrastive loss into unified contrastive loss, accommodating arbitrary positives and negatives within a unified pair-wise optimization framework. Chen et al. [49] aligned input examples and classifiers within a shared representation space, performing contrastive learning between them to augment classification efficiency.

This multifaceted foundation underpins our study, introducing a novel approach that applies supervised cross MoCo framework (SuCroMoCo) for financial and prototypical examples alignment, culminating in augmented performance across financial sentiment analysis benchmarks. Moreover, contrastive learning manifests potential beyond the domain of augmentation, extending into the alignment of features across disparate domains to create a unified representation space. For instance, Wang et al. [50] introduced bidirectional matching for bilingual sentence representation, capitalizing on two momentum encoders and amalgamating two contrastive losses. The work of Tan et al. [51] stands as a testament to this, as they employed supervised contrastive learning with a memory bank [52] to enhance the classification efficiency of PLMs operating across distinct textual domains that share common categories. Lastly, Wang et al. [53] engineered a cross-domain contrastive self-supervised learning framework, effectively harmonizing distinct visual features characterized by varying styles, thus reducing domain discrepancies between training and testing sets. This paradigm provides us with the means to utilize examples from diverse sources as prototypes.

## 3. Methodology

As depicted in Fig. 2, our proposed SuCroMoCo extends the self-supervised MoCo framework to a fully supervised learning context. Importantly, we utilize prototypical examples rather than self-augmentation for the contrasting objective, aiming to align financial representations with prototype representations that are categorized based on sentiment. In this section, we initially introduce the fundamental concepts of the MoCo framework, and explore its implementation in both self-supervised and supervised scenarios. Subsequently, we will detail the setup of our study and delve into the specifics of the SuCroMoCo framework.

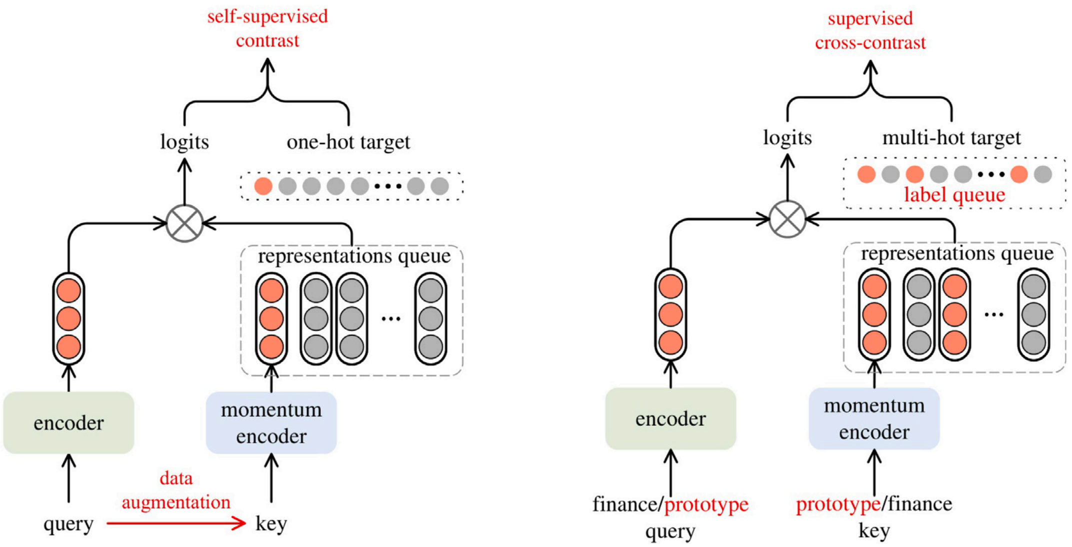
### 3.1. MoCo

MoCo adopts a distinctive perspective on contrastive learning, treating the amalgamation of examples and their corresponding augmentations as keys in a dynamic dictionary. This dictionary is effectively maintained through the use of a queue, for the purpose of dictionary look-up. Given a set of arbitrary augmented labeled text examples $\mathcal{D} = \left\{(x_i, y_i)\right\}$, let $i \in I \equiv \{1, \ldots, 2N\}$ be the index of the samples, and let $j(i)$ be the index of the other augmented examples originating from the same source one. The self-supervised contrastive loss can be denoted as follows:

$$\mathcal{L}^{self} = -\sum_{i\in I} \log \frac{\exp\left(sim\left(z_i, z_{j(i)}\right)/\tau\right)}{\sum_{a\in A(i)} \exp\left(sim\left(z_i, z_a\right)/\tau\right)} \tag{1}$$

Here, $z_i = E_{\theta_q}(x_i)$ is the encoded representation of the input example, where $E_{\theta_q}$ corresponds to the query encoder with parameters $\theta_q$. The representations $z_{j(i)}$ and $\left\{z_a\right\}$ collectively form the keys queue. This queue is maintained by a key encoder $E_{\theta_k}$, which undergoes updates using a momentum-based moving average of the query encoder. $A(i) \equiv I \setminus \{i\}$ is the set of all the indices except the query index $i$. In this context, the $j(i)$th augmented example is called the positive, and the other remaining $2(N-1)$ examples are called the negatives. The underlying objective of MoCo is to classify $z_i$ as the positive pair of $z_{j(i)}$ among the entire set of representations. The function $sim()$ represents

[1] Interestingly, the most consistent financial PLMs across different tasks have been shown to be those retaining the original vocabulary of general domain models [20,42].



(a) MoCo (b) SuCroMoCo

**Fig. 2.** Comparison of MoCo and SuCroMoCo. Unlike MoCo, which generates keys through data augmentation, SuCroMoCo conducts cross-contrast using prototypical examples as keys for financial queries and financial keys for prototypical queries. Positive samples encompass those with the same sentiment class. The length of the label queue matches that of the keys queue.

the cosine similarity function, and $\tau \in \mathcal{R}^+$ stands as the temperature hyperparameter. Notably, empirical observations have highlighted that modulating the temperature can significantly enhance performance.

Moving into a supervised context, examples within the same category can serve as positives for contrastive learning. Consequently, multiple potential positives emerge within the keys queue. For an input query $x_i$ and its label $y_i$, the set of indices of potential positives can be expressed as $P(i) \equiv \{p \in A(i) : y_p = y_i\}$. The self-supervised contrastive loss can be extended to formulate a supervised contrastive loss, as shown below:

$$\mathcal{L}^{sup} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(sim\left(z_i, z_p\right)/\tau\right)}{\sum_{a\in A(i)} \exp\left(sim\left(z_i, z_a\right)/\tau\right)} \tag{2}$$

Throughout the training process, the encoded representations of the current mini-batch are enqueued, while the oldest entries are dequeued. This utilization of queue effectively decouples the size of the dictionary from the mini-batch size, thereby allowing the dictionary to be sizeable. Importantly, in supervised context, a corresponding label queue is necessary and updated simultaneously to provide label information.

In order to handle the challenge of updating the key encoder with a large dictionary via back-propagation, MoCo employs a momentum update mechanism, as follows:

$$\theta_k \leftarrow m\theta_k + (1-m)\,\theta_q \tag{3}$$

In Eq. (3), $m \in [0, 1]$ embodies the momentum coefficient. Notably, a higher momentum coefficient engenders a slower and more gradual evolution of the key encoder. The momentum update contributes to minimizing discrepancies between keys across diverse mini-batches, resulting in enhanced consistency and improved overall performance.

### 3.2. SuCroMoCo

To align financial text with prototypical examples in representation space, we generalize supervised MoCo to SuCroMoCo to contrast financial samples against prototypical examples by treating the prototypical examples as augmentations.

In detail, we denote our target financial dataset as $\mathcal{D}^{Fin} = \{x_i^{Fin}, y_i^{Fin}\}$, with $i \in I^{Fin} \equiv \{1, \ldots, N\}$, and the dataset of prototypical examples as $\mathcal{D}^{Pro} = \{(x_k^{Pro}, y_k^{Pro})\}$, where $k \in I^{Pro} \equiv \{1, \ldots, M\}$. Importantly, both $\{y_i^{Fin}\}$ and $\{y_k^{Pro}\}$ are encompassed within the same pre-defined label space, which delineates the categories. Through maintaining a representation queue and a corresponding label queue for all the prototypical examples, an input financial instance can contrast against multiple prototypes at the same time. To this end, for a input financial instance $x_i$, the set of indices of positives is denoted as $P^F(i) \equiv \{p \in A^{Fin}(i) : y_p^{Fin} = y_i^{Fin}\} \cup \{p \in I^{Pro} : y_p^{Pro} = y_i^{Fin}\}$. The supervised loss of contrasting financial representations against prototypical representations can be expressed as below:

$$\mathcal{L}^{SuCroMoCo}_{Fin} = \sum_{i\in I^{Fin}} \frac{-1}{|P^{Fin}(i)|} \sum_{p\in P^{Fin}(i)} \log \frac{\exp\left(sim\left(z_i, z_p\right)/\tau\right)}{\sum_{a\in A^{Fin}(i)} \exp\left(sim\left(z_i, z_a\right)/\tau\right)} \tag{4}$$

During the training procedure, the input mini-batch of financial queries will propagate gradient to train query encoder, forcing it to gradually pull financial representations closer to prototypical examples. Moreover, we argue that the prototypical representations should also be pulled closer to financial instances for a better clustering. Therefore, we also consider a prototypical text as a input query, and contrast it against all of the financial instances, which now are functioned as keys. The $\mathcal{L}^{SuCroMoCo}_{Pro}$ is computed similarly to Eq. (4) by setting $P^{Pro}(i) \equiv \{p \in A^{Pro}(i) : y_p^{Fin} = y_i^{Fin}\} \cup \{p \in I^{Pro} : y_p^{Pro} = y_i^{Fin}\}$. Then, we combine $\mathcal{L}^{SuCroMoCo}_{Fin}$ with $\mathcal{L}^{SuCroMoCo}_{Pro}$ to derive the supervised cross-momentum loss as follows:

$$\mathcal{L}^{SuCroMoCo} = \mathcal{L}^{SuCroMoCo}_{Fin} + \mathcal{L}^{SuCroMoCo}_{Pro} \tag{5}$$

In particular, we maintain two separate key queues and their corresponding label queues for all of the financial and prototypical representations. The key encoder undergoes a momentum-based update after each batch step's bidirectional contrast, generating new representations both queues. Additionally, a classifier $C_\phi$ takes the representations of both financial and prototypical examples as input and predicts the category $y\prime$, denoted as $y\prime = C_\phi(z)$. The final loss function is the combination of the supervised cross-entropy loss and the SuCroMoCo loss, performed in a fully supervised condition:

$$\mathcal{L} = \mathcal{L}^{CE}\left(E_{\theta_q}, C_\phi\right) + \mathcal{L}^{SuCroMoCo}\left(E_{\theta_q}\right) \tag{6}$$

The pseudo code of our proposed SuCroMoCo method is outlined in Algorithm 1.

**Algorithm 1:** Pseudo code of SuCroMoCo in a PyTorch-like style.

```
# m: momentum coefficient
# t: temperature
# BS: batch size
# d: representation dimension
# E_q, E_k: encoder networks for query and key
# que_fin: financial representation queue of N
    keys (N*d)
# que_pro: prototypical representation queue of
    M keys (M*d)
# labels_fin: financial label queue of N keys (N
    )
# labels_pro: prototypical label queue of M keys
     (M)

# initialize
E_k.params = E_q.params
for (x_fin,l_fin,x_pro,l_pro)
    in zip(fin_loader, pro_loader):

    # financial queries: (BS*d)
    q_fin = E_q.forward(x_fin)
    # financial logits: (BS*M)
    logits_fin = cos_sim(q_fin.view(N, 1, d),
                         que_pro.view(1, M, d),
    dim=-1)

    # prototypical queries (BS*d)
    q_pro = E_q.forward(x_pro)
    # prototypical logits: (BS*N)
    logits_pro = cos_sim(q_pro.view(M, 1, d),
                         que_fin.view(1, N, d),
    dim=-1)

    target_fin = (l_fin[:, None] == labels_pro[
    None, :])
    target_pro = (l_pro[:, None] == labels_fin[
    None, :])

    loss_fin = SuCroMoCo(logits_fin/t,
                        target_fin.float())
    loss_pro = SuCroMoCo(logits_pro/t,
                        target_pro.float())
    loss = loss_fin+loss_pro
    loss.backward()

    update(E_q.params)
    E_k.params = m*E_k.params+(1-m)*E_q.params

    k_fin = E_k.forward(x_fin).detach()
    k_pro = E_k.forward(x_pro).detach()

    enqueue_dequeue(que_fin, k_fin)
    enqueue_dequeue(que_pro, k_pro)
    enqueue_dequeue(labels_fin, l_fin)
    enqueue_dequeue(labels_pro, l_pro)
```

## 4. Experiments

In this section, we present an empirical evaluation of the proposed SuCroMoCo framework using three benchmark financial datasets. The primary objective is to assess the performance of SuCroMoCo in comparison to various financial PLMs and recently introduced large language models (LLMs). Additionally, we extend the comparison to include a state-of-the-art contrastive learning-based sentiment classification approach and a cross-domain sentiment classification method. BERT and RoBERTa serve as the foundational models for all the PLM-based approaches, including SuCroMoCo. In addition to the comparison experiments, we conduct visualization and ablation studies to gain deeper insights into the functioning and effectiveness of the proposed SuCroMoCo framework.

### *4.1. Datasets*

**Benchmark Datasets**. For our evaluation, we have selected three benchmark datasets of sentiment analysis in the financial domain. These datasets include both social media data and financial news data, ensuring a diverse and robust evaluation of our proposed SuCroMoCo framework. The datasets are as follows:

**StockSen** [22]: This dataset consists of 20,675 financial tweets sourced from the StockTwits platform, spanning the period from June to August 2019. Each tweet has been annotated with either a positive or negative sentiment label. The dataset is divided into a training set with 14,457 instances and a development set with 6218 instances, which we treat as the test set.

**TweetFinSent** [23]: This dataset is constructed from financial tweets and encompasses 2113 sentences that have been annotated with positive, neutral, and negative polarities. After extracting the tweets from the internet, the dataset was ultimately composed of 821 instances for the training set and 697 instances for the test set. Some tweets were either removed or their privacy settings were altered during the data collection process.

**FinTextSen** [54,55]: Initially introduced in SemEval-2017 Task 5 [54], this dataset focuses on fine-grained sentiment analysis of financial microblogs. Post-processing, it comprises 2486 microblog messages sourced from Twitter and StockTwits in March 2016. Each instance includes the message, a cashtag, and a continuous sentiment score ranging from −1 (negative) to 1 (positive). Daudert et al. [55] grouped the scores into a 3-class annotation (Positive, Negative, and Neutral). For consistency, we use the categorical version of this dataset in our study.

**Financial PhraseBank (FPB)** [21]: This dataset offers sentiment-labeled sentences extracted from financial news articles. Annotators categorized the sentences into positive, neutral, and negative sentiments. The dataset is provided in multiple subsets based on annotator agreement:

- FPB-50%: 4846 instances with at least 50% agreement.
- FPB-66%: 4217 instances with at least 66% agreement.
- FPB-75%: 3453 instances with at least 75% agreement.
- FPB-100%: 2264 instances with 100% agreement.

Results reported in [20] have convincingly highlighted have showcased that PLMs on the FPB-100% subset obtain an almost perfect performance. Consequently, rather than conducting our evaluation on this specific subset, we employed the instances from FPB-100% as prototypes to train the SuCroMoCo framework, and thus we focused just on the FPB-50% subset as a test set. Additionally, we took care to exclude any examples within the FPB-50% subset that overlapped with the FPB-100% subset. Following this separation process, the FPB-50% subset comprised a total of **2797** instances available for our experiments.

**Prototypical Examples**. We selected prototypical examples from the datasets that share the same categorical labels as those present in our benchmark datasets. This strategy ensured that the prototypical examples were relevant and representative of the sentiment categories within our evaluation.

We opted to use instances from the **Stanford Sentiment TreeBank binary (SST-2)** [56] dataset as prototypes for the StockSen dataset. The SST-2 dataset comprises 11,855 sentences that are extracted from movie reviews, and each sentence is annotated with either a positive or negative label, making it directly compatible with the sentiment labels in the StockSen dataset. This dataset has been widely employed in binary classification experiments since the early 2010s, and it has served as a benchmark for assessing the performance of PLMs, which have consistently demonstrated impressive results on it [57]. Therefore, it is a good candidate to provide prototypical examples for our binary dataset.

Given that both the TweetFinSent and FPB-50% datasets involve instances annotated with three sentiment polarities, and considering that PLMs could achieve remarkable performance on the FPB-100% subset, we decided to use the instances from **Financial PhraseBank 100% (FPB-100%)** as prototypes for both the TweetFinSent and FPB-50% datasets. Moreover, to ensure the independence of our evaluation, we excluded any overlapping instances between these subsets.

### 4.2. Baselines

To ensure fairness in our comparative analysis, we employ the same set of prototypical examples to train not only our SuCroMoCo model but also the baseline models. The baseline models considered for evaluation in our study are as follows:

- **BERT** [9] and **RoBERTa** [10]. Given the foundational role of BERT and RoBERTa in our approach and other baseline models, we adopt them as baseline models and present the fine-tuning outcomes for both of them. The fine-tuning process entails adding a linear layer head on top of the models and employing cross-entropy loss with back-propagation.
- **FinBERT**[2] [17]: FinBERT is a BERT-based model that has been pre-trained on financial communication data, including corporate reports, earnings call transcripts, and analyst reports. Two variants of FinBERT -FinBERT-BaseVocab and FinBERT-FinVocab- are openly accessible, and they differ for their vocabulary: the former shares the general domain vocabulary with the original BERT model, while the latter has a vocabulary specific for financial texts. In our study, we utilize FinBERT-BaseVocab as it was proved to have a better performance on sentiment analysis tasks [20].
- **FLANG**[3] [19]: FLANG is a collection of large language models specifically designed for financial natural language processing tasks. FLANG models undergo domain-specific pre-training by preferentially masking words and phrases according to financial dictionaries. We fine-tune both FLANG$_{\text{BERT}}$ and FLANG$_{\text{RoBERTa}}$ on our benchmark datasets for comparison.

Specifically, the evaluation outcomes of the aforementioned models are presented separately for two scenarios: fine-tuning exclusively on financial datasets, and fine-tuning jointly on a combination of prototypical examples and financial datasets.

- **DualCL**[4] [49]: DualCL extends supervised contrastive learning by incorporating label words, e.g. “positive” and “negative”, directly into sentences as classifiers. It simultaneously learns both sentence and classifier representations, treating the classifier representations as a form of augmentation. The primary objective is to leverage supervised contrastive learning to minimize the distance between a sentence representation and the correct classifier representation, while maximizing the separation from the representation of the opposite classifier. This approach enhances the model’s capability to acquire representations that are highly discriminative and closely aligned with the provided labels. To ensure a fair comparison, we also train the DualCL model on a combined dataset comprising prototypical examples and financial data.
- **AdSPT** [58]: AdSPT is a cross-domain sentiment classification method that leverages soft prompts to learn distinct vectors for various domains. It achieves domain-invariant representations through adversarial training. In our experiments, we re-implement the AdSPT model based on the original paper’s methodology. We perform experiments on our benchmark datasets using the set of prototypical examples as the model’s source domains. Unlike the original configuration, we fine-tune the classifier of the AdSPT model using both the financial and prototypical examples. This modification ensures a fair comparison between AdSPT and other models.

Furthermore, we expanded our comparative analysis to incorporate two recent open-source *Large Language Models* (LLMs) [59]: LLaMA 2 [60] and FinMA [61]:

- **LLaMA 2** [60]: LLaMA 2, an open-source pre-trained language model, represents an upgraded iteration of LLaMA [62]. Notable improvements include a larger training corpus, extended context length, and alignment with human preferences and safety standards. Our experiments leverage the `LLaMA-2-7b-hf`[5] model.
- **FinMA** [61]: FinMA stands out as the first open-source financial large language model, derived from fine-tuning LLaMA [62] with a curated financial instructions dataset. Our study employs the `FinMA-7B-NLP`[6] model, trained on NLP tasks specific to finance from the PIXIU dataset [61].

Initially, we fine-tuned LLaMA 2 and FinMA in a supervised manner using Low-Rank Adaptation (LoRA) [63]. In this scenario, LLMs undergo fine-tuning alongside a classification head, and updates are propagated through back-propagation.

In addition, LLMs have introduced a new paradigm in NLP that eliminates the need for model finetuning. It has been suggested that for models exceeding approximately 6 billion parameters, a “phase shift” occurs in the weights, resulting in the emergence of outliers in key features responsible for significantly enhanced performance and in-context learning. This capacity allows models to perform tasks they were not explicitly trained for [64]. Large Language Models, therefore, are typically employed via *prompting* in zero-shot or few-shot learning scenarios. Given their lack of dependency on fine-tuning, they present an attractive option for NLP practitioners in the financial field. Consequently, we opted to include the prompting mechanism in our comparative evaluation. Specifically, we present the results of prompting ChatGPT.[7]

ChatGPT, an advanced language model developed by OpenAI, is constructed upon the GPT-3.5 architecture with an extensive parameter count totaling 175 billion. It underwent fine-tuning using reinforcement learning with human feedback [65]. Designed primarily for generating human-like text based on input, ChatGPT has been trained on a diverse range of internet text sources. In this study, we experiment with ChatGPT (`gpt-3.5-turbo`) for the FSA task using the OpenAI API.[8]

### 4.3. Implementation details

**Supervised Fine-tuning**: We adopt `bert-base-uncased` and `roberta-base` models as the basis of DualCL, AdSPT, and SuCroMoCo. To enhance accessibility, we load Language Models (LLMs) in half-precision floating-point format (`fp16`) [66] and implement gradient accumulation [67] during the fine-tuning of LLaMA 2 and FinMA. For the LoRA configuration, the rank is set to 8, resulting in

[2] https://github.com/yya518/FinBERT
[3] https://github.com/SALT-NLP/FLANG
[4] https://github.com/hiyouga/Dual-Contrastive-Learning
[5] https://huggingface.co/meta-llama/Llama-2-7b-hf
[6] https://huggingface.co/ChanceFocus/finma-7b-nlp
[7] https://chat.openai.com/chat
[8] https://platform.openai.com/docs/introduction

**Table 1**
The optimal combination of temperature ($\tau$) and momentum ($m$) obtained after grid searching.

| | BERT | | RoBERTa | |
|---|---|---|---|---|
| | $\tau$ | $m$ | $\tau$ | $m$ |
| StockSen | 0.04 | 0.999 | 0.06 | 0.99 |
| TweetFinSent | 0.03 | 0.999 | 0.08 | 0.99 |
| FinTextSen | 0.07 | 0.99 | 0.08 | 0.99 |
| FPB-50% | 0.1 | 0.999 | 0.05 | 0.999 |

approximately 4 million trainable parameters for both LLaMA 2 and FinMA.

The evaluation process is performed at the conclusion of each epoch, with early stopping integrated to capture the optimal results. Our early stopping strategy is set with a patience value of 3 epochs. We run every approaches 5 times on the benchmarks that have official train-test split, e.g., StockSen and TweetFinSent and we take the average score. For the FinTextSen and FPB-50% benchmarks, 5-fold cross validation is implemented for a robust evaluation. The average accuracy and macro-F1 scores are reported for comparison.

To fine-tune the hyperparameters effectively, we conduct grid searches across different datasets to identify the optimal combination of the temperature ($\tau$) and momentum ($m$) parameters for various backbone models. The selected hyperparameter combinations are detailed in Table 1. The analysis of the impact of these two hyperparameters will be presented in Section 5.4.

The experiments have been run on two NVIDIA GeForce RTX 3090 GPUs, boasting a total of 48 GB memory.

**Prompting LLMs**: We conduct experiments on the three LLMs in both zero-shot and few-shot scenarios. The input prompt is structured following the instruction template outlined in Xie et al. [61], which is applied consistently across models. The prompt is designed as follows:

- Analyze the sentiment of this statement extracted from [Dataset]. You must provide your answer as either negative or positive.

The [Dataset] placeholder indicates the source of the input sentences. In the StockSen, TweetFinSent, and FinTextSen datasets, we use the phrase "a financial tweet" to fill in the [Dataset] placeholder. For the FPB dataset, we replace the [Dataset] placeholder with the phrase "a financial news article". In a zero-shot scenario, the input instance is simply appended to the end of the prompt. Conversely, in few-shot scenarios, we randomly select three samples from each category. These selected samples, along with their corresponding labels, are then incorporated into the prompt.

## 5. Experimental results and discussion

### 5.1. Supervised fine-tuning

The comparison results between the proposed SuCroMoCo and other supervised approaches are presented in Table 2. Leveraging BERT as the foundation, SuCroMoCo consistently achieves the highest accuracy and macro-F1 scores on StockSen, TweetFinSent and FPB-50% datasets, attesting to its robust performance. With RoBERTa as the foundation, SuCroMoCo excels in StockSen and FPB-50%, maintaining strong competitiveness across datasets. Notably, SuCroMoCo exhibits a trade-off between accuracy and macro-F1 scores on the FinTextSen dataset when compared to $\text{BERT}_{\text{joint}}$ and $\text{RoBERTa}_{\text{joint}}$, showcasing its proficiency in handling class imbalances. Despite a slight underperformance against DualCL on the TweetFinSent dataset, SuCroMoCo's overall performance remains competitive. The observed instability in DualCL's performance, evident in decreased accuracy on the FPB-50% dataset for both BERT and RoBERTa foundations compared to fine-tuning on vanilla models, showcases that aligning financial text representations with prototypical counterparts is more stable than aligning financial text representations with classifier word representations. This result highlights the broad applicability and efficacy of the proposed SuCroMoCo approach.

DualCL contrasts different inputs against same classifier words, notably 'positive' and 'negative'. While these words are inserted into different sentences, their inherent meanings tend to rigidly fix their representation vectors within the broader representation space. Moreover, the sentiment of financial texts is not always directly linked to terms like 'positive' and 'negative'. Furthermore, the PLMs might lack the domain-specific knowledge insights required to fully comprehend the intricate meaning of financial texts. As a result, aligning them with classifier representations can be challenging, leading to the observed performance fluctuations in DualCL. In contrast, our SuCroMoCo approach adopts a more resilient strategy. By contrasting a single input against multiple instances within the same sentiment category, we mitigate the potential for misalignment between representations, which contributes to the enhanced performance.

The less favorable outcomes of AdSPT underscore the limitations of relying on adversarial generative learning to align financial representations with prototypical examples. Throughout our experimentation, we observed that the generator struggled to effectively transform financial representations into prototypical representations that could deceive the discriminator. This difficulty in achieving meaningful alignment ultimately contributed to the observed decrease in performance when compared to other methods.

Supervised fine-tuning of LLMs demonstrates instability. Although LLaMA 2 and FinMA namely attain the highest scores on StockSen and FPB datasets, they lag behind other approaches on the remaining three datasets. The substantial improvement observed of FinMA on FPB dataset is probably due to label leaking, as the model was exposed to the FPB dataset during the instruction tuning stage [61] (nonetheless, we still decided to report the scores of FinMA for consistency in the evaluation results).

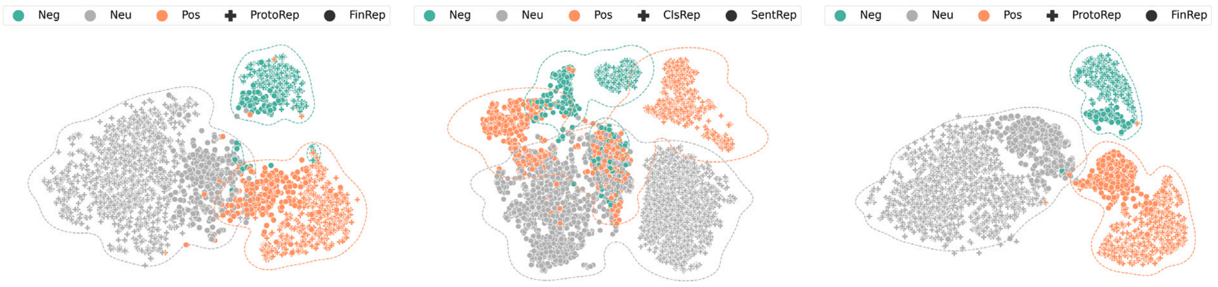
In addition to the unstable performance, the time cost of fine-tuning LLMs is very significant. Table 3 illustrates the time cost per epoch for different models training on the same dataset. It takes almost 10 times longer to fine-tune an LLM than it does to fine-tune the proposed model. During the experiment, we observed that this is mainly due to the quantization process. The extended time consumption and larger memory footprint make fine-tuning LLMs a costly and lengthy process.

A notable observation is that simultaneously fine-tuning models on a combination of financial and prototypical examples can yield performance enhancements for both the original and domain-specific PLMs. This fine-tuning strategy is sometimes even more effective than domain adaptive pre-training on Financial FSA tasks. specifically, BERTjoint achieves higher accuracy and macro-F1 scores on TweetFinSent and FPB-50% datasets compared to FinBERT and $\text{FLANG}_{\text{BERT}}$. Similarly, $\text{RoBERTa}_{\text{joint}}$ surpasses $\text{FLANG}_{\text{RoBERTa}}$ in terms of performance across all benchmark datasets. These enhancements also occur on LLMs on most datasets. This phenomenon can be attributed to the fact that supervised fine-tuning inherently involves representation alignment. To facilitate accurate classification, supervised fine-tuning optimizes the parameters of the PLM to ensure that inputs with the same sentiment are closely represented in the embedding space. Our SuCroMoCo takes this alignment a step further by leveraging supervised cross-contrast to effectively align representations, ultimately resulting in superior classification outcomes.

To explore the representation acquisition processes of various methods, we conducted a visualization analysis of the learned representation distribution. This investigation involved capturing the representations from the converged training epoch of (i) $\text{BERT}_{\text{joint}}$, (ii) $\text{DualCL}_{\text{BERT}}$, and (iii) $\text{SuCroMoCo}_{\text{BERT}}$, all trained on the TweetFinSent dataset. Subsequently, we applied dimensionality reduction techniques to generate scatter plots visually depicting the distribution of representations. The outcomes of this visualization are presented in Fig. 3

**Table 2**
Comparison results of supervised fine-tuning approaches are presented in three sections. The upper part displays results for BERT-based models, the intermediate part presents results for RoBERTa-based models, and the bottom part showcases results for supervised fine-tuned LLMs. Reported values are averages with corresponding standard deviations shown in subscripts, obtained from either 5 independent runs or 5-fold cross-validation. Performance scores highlighted in red signify the best results among models sharing the same foundation PLM. Scores marked in **bold** and red denote the overall best performance across approaches based on different foundation PLMs.

| Models | StockSen | | TweetFinSent | | FinTextSen | | FPB-50% | |
|---|---|---|---|---|---|---|---|---|
| | Acc(%) | Macro-F1(%) | Acc(%) | Macro-F1(%) | Acc(%) | Macro-F1(%) | Acc(%) | Macro-F1(%) |
| BERT | $78.22_{0.24}$ | $72.69_{0.38}$ | $56.01_{0.91}$ | $48.57_{1.56}$ | $83.57_{2.33}$ | $60.92_{5.35}$ | $77.39_{0.72}$ | $76.70_{0.84}$ |
| $\text{BERT}_{\text{joint}}$ | $79.43_{0.96}$ | $73.02_{0.57}$ | $57.59_{1.46}$ | $52.23_{1.23}$ | $84.17_{1.68}$ | $62.86_{2.72}$ | $79.04_{2.36}$ | $78.23_{2.29}$ |
| FinBERT | $79.69_{0.43}$ | $73.19_{0.33}$ | $55.24_{0.92}$ | $47.80_{1.59}$ | $86.14_{1.99}$ | $63.70_{4.47}$ | $76.93_{0.70}$ | $76.14_{0.75}$ |
| $\text{FinBERT}_{\text{joint}}$ | $79.69_{0.93}$ | $73.08_{0.55}$ | $57.33_{2.24}$ | $51.37_{1.82}$ | $86.06_{0.81}$ | $64.51_{2.72}$ | $79.11_{1.09}$ | $78.36_{1.79}$ |
| $\text{FLANG}_{\text{BERT}}$ | $79.67_{0.59}$ | $73.41_{0.22}$ | $54.32_{1.09}$ | $46.39_{1.44}$ | $83.57_{1.46}$ | $61.80_{3.56}$ | $77.14_{1.33}$ | $76.17_{1.35}$ |
| $\text{FLANG}_{\text{BERT}_{\text{joint}}}$ | $79.57_{0.93}$ | $73.80_{0.59}$ | $58.31_{1.07}$ | $52.77_{1.10}$ | $84.70_{1.62}$ | $63.93_{3.34}$ | $78.36_{0.93}$ | $77.57_{1.04}$ |
| $\text{DualCL}_{\text{BERT}}$ | $79.25_{0.22}$ | $73.11_{0.24}$ | $55.78_{0.73}$ | $49.21_{1.04}$ | $82.89_{1.97}$ | $59.78_{3.85}$ | $77.07_{2.58}$ | $76.37_{2.88}$ |
| $\text{AdSPT}_{\text{BERT}}$ | $67.57_{2.63}$ | $60.71_{1.75}$ | $52.94_{2.61}$ | $42.75_{2.55}$ | $78.33_{5.35}$ | $53.72_{3.19}$ | $75.11_{1.67}$ | $71.00_{2.67}$ |
| $\text{SuCroMoCo}_{\text{BERT}}$ | $80.42_{0.26}$ | $74.15_{0.30}$ | $59.25_{0.39}$ | $54.45_{0.79}$ | $84.70_{1.46}$ | $65.02_{3.87}$ | $79.40_{2.32}$ | $78.43_{2.78}$ |
| RoBERTa | $81.18_{0.70}$ | $76.02_{0.51}$ | $57.88_{1.56}$ | $53.56_{2.17}$ | $87.27_{1.93}$ | $63.17_{5.28}$ | $79.64_{0.73}$ | $79.19_{0.67}$ |
| $\text{RoBERTa}_{\text{joint}}$ | $82.10_{0.78}$ | $75.82_{1.28}$ | $60.89_{1.55}$ | $56.74_{1.02}$ | $\mathbf{87.87}_{1.44}$ | $65.42_{4.46}$ | $80.27_{0.73}$ | $79.63_{0.61}$ |
| $\text{FLANG}_{\text{RoBERTa}}$ | $81.01_{0.48}$ | $75.22_{0.17}$ | $59.37_{0.86}$ | $51.61_{0.98}$ | $85.30_{2.21}$ | $56.43_{1.63}$ | $76.39_{1.82}$ | $75.82_{1.55}$ |
| $\text{FLANG}_{\text{RoBERTa}_{\text{joint}}}$ | $81.35_{0.34}$ | $75.22_{0.17}$ | $60.04_{1.05}$ | $56.53_{0.39}$ | $85.50_{1.50}$ | $62.65_{4.08}$ | $78.57_{1.63}$ | $77.77_{1.59}$ |
| $\text{DualCL}_{\text{RoBERTa}}$ | $82.24_{0.21}$ | $76.45_{0.16}$ | $\mathbf{61.22}_{0.75}$ | $\mathbf{58.01}_{0.90}$ | $87.43_{1.72}$ | $62.02_{4.37}$ | $79.11_{1.24}$ | $78.76_{1.52}$ |
| $\text{AdSPT}_{\text{RoBERTa}}$ | $69.87_{1.63}$ | $61.23_{1.45}$ | $54.86_{2.50}$ | $48.38_{1.83}$ | $80.72_{3.04}$ | $55.68_{2.17}$ | $78.07_{3.06}$ | $76.04_{3.36}$ |
| $\text{SuCroMoCo}_{\text{RoBERTa}}$ | $82.59_{0.28}$ | $77.02_{0.37}$ | $61.00_{1.03}$ | $57.47_{0.56}$ | $87.75_{1.21}$ | $\mathbf{66.50}_{5.47}$ | $80.84_{1.03}$ | $80.34_{1.12}$ |
| LLaMA 2 | $\mathbf{84.03}_{0.44}$ | $\mathbf{79.84}_{0.36}$ | $52.46_{0.14}$ | $42.92_{1.64}$ | $72.97_{2.59}$ | $48.03_{1.93}$ | $67.14_{2.19}$ | $64.06_{3.30}$ |
| LLaMA $2_{\text{joint}}$ | $83.87_{0.18}$ | $79.44_{0.28}$ | $51.36_{0.84}$ | $42.39_{0.52}$ | $85.34_{1.85}$ | $57.66_{3.03}$ | $77.57_{0.57}$ | $76.73_{1.04}$ |
| FinMA | $83.38_{0.24}$ | $78.48_{0.16}$ | $51.51_{1.53}$ | $40.31_{1.57}$ | $83.89_{1.06}$ | $55.22_{0.95}$ | $90.46_{0.79}$ | $90.17_{1.02}$ |
| $\text{FinMA}_{\text{joint}}$ | $82.29_{0.35}$ | $75.28_{2.88}$ | $53.52_{1.46}$ | $46.06_{1.66}$ | $86.75_{1.20}$ | $60.99_{2.14}$ | $\mathbf{91.68}_{0.96}$ | $\mathbf{91.63}_{0.91}$ |



(a) $\text{BERT}_{\text{joint}}$ (b) $\text{DualCL}_{\text{BERT}}$ (c) $\text{SuCroMoCo}_{\text{BERT}}$

**Fig. 3.** The t-SNE plots depict the learned representations on the TweetFinSent dataset for three different approaches: $\text{BERT}_{\text{joint}}$, $\text{DualCL}_{\text{BERT}}$, and $\text{SuCroMoCo}_{\text{BERT}}$. In sub-figures (a) and (c), circular markers denote financial representations (FinRep), and cross markers signify prototypical representations (ProtoRep). In sub-figure (b), circular markers denote sentence representations (SentRep) encompassing both financial and prototypical sentences, while plus markers represent classifier representations (ClsRep). Different colors have been applied to represent various sentiment categories.

**Table 3**
Average training time (seconds) per epoch for different models on the TweetFinSent dataset.

| Models | Time |
|---|---|
| RoBERTa | 18.89 |
| $\text{DualCL}_{\text{RoBERTa}}$ | 19.34 |
| $\text{SuCroMoCo}_{\text{RoBERTa}}$ | 26.83 |
| LLaMA 2 | 232.99 |
| FinMA | 234.01 |

In Fig. 3(a), it is observed that both financial and prototypical representations tend to cluster into three distinguishable groups after supervised fine-tuning. Notably, prototypical representations corresponding to distinct sentiment polarities exhibit clear separation in the space, while financial representations tend to aggregate more closely, although some still maintain alignment with their specific sentiment categories.

In Fig. 3(b), the learned sentence representations of $\text{DualCL}_{\text{BERT}}$ exhibit a degree of proximity among themselves, with some representations being indistinguishable based on sentiment categories. However, the representations stemming from classifiers, which are aligned with different sentiment categories, appear to be more distant from one another. It is crucial to highlight that the financial representations do not appear to effectively converge towards the classifier representations associated with corresponding sentiment categories. This could potentially contribute to the observed decrease in performance in the case of $\text{DualCL}_{\text{BERT}}$.

In contrast, the financial representations acquired through $\text{SuCroMoCo}_{\text{BERT}}$ demonstrate a different behavior. These representations exhibit a tendency to cluster closer to the prototypical representations that share the same sentiment category. Moreover, they deliberately maintain distance from representations associated with different sentiment classifications. This observation suggests that the proposed SuCroMoCo effectively aligns representations, which likely contributes to the performance enhancement witnessed in our experiments.

### 5.2. Prompting LLMs

We conducted an evaluation of LLMs using 0/3-shot prompting on four datasets. In this assessment, a response generated by the LLMs containing any of the words ‘positive,’ ‘neutral,’ or ‘negative’ is considered as a valid classification. The outcomes are presented in Table 4.

In the 0-shot scenario, ChatGPT struggles to provide valid predictions for all instances across the three social media datasets. Similarly,

**Table 4**
The classification results for the four LLMs using 0/3-shot prompting.

| Dataset | Size | Model | Valid predictions | Overall Acc(%) | Valid Acc(%) | Valid Macro-F1(%) |
|---|---|---|---|---|---|---|
| StockSen | 6218 | LLaMA $2_{0-\text{shot}}$ | 936 | 8.14 | 54.06 | 51.72 |
| | | LLaMA $2_{3-\text{shot}}$ | 324 | 3.07 | 58.95 | 50.93 |
| | | $\text{FinMA}_{0-\text{shot}}$ | 2282 | 29.85 | 81.33 | 76.39 |
| | | $\text{FinMA}_{3-\text{shot}}$ | 717 | 9.38 | 81.31 | 80.92 |
| | | $\text{ChatGPT}_{0-\text{shot}}$ | 5330 | 68.91 | 80.39 | 76.35 |
| | | $\text{ChatGPT}_{3-\text{shot}}$ | 6156 | 49.26 | 49.76 | 30.33 |
| | | $\text{SuCroMoCo}_{\text{BERT}}$ | 6218 | – | $80.42_{0.26}$ | $74.15_{0.30}$ |
| | | $\text{SuCroMoCo}_{\text{RoBERTa}}$ | 6218 | – | $82.59_{0.28}$ | $77.02_{0.37}$ |
| TweetFinSent | 697 | LLaMA $2_{0-\text{shot}}$ | 76 | 4.45 | 40.79 | 37.19 |
| | | LLaMA $2_{3-\text{shot}}$ | 7 | 0.29 | 29.57 | 20.63 |
| | | $\text{FinMA}_{0-\text{shot}}$ | 150 | 11.91 | 55.33 | 54.62 |
| | | $\text{FinMA}_{3-\text{shot}}$ | 125 | 9.75 | 54.4 | 49.11 |
| | | $\text{ChatGPT}_{0-\text{shot}}$ | 692 | 58.39 | 58.82 | 56.01 |
| | | $\text{ChatGPT}_{3-\text{shot}}$ | 697 | – | 54.23 | 51.51 |
| | | $\text{SuCroMoCo}_{\text{BERT}}$ | 697 | – | $59.25_{0.39}$ | $54.45_{0.79}$ |
| | | $\text{SuCroMoCo}_{\text{RoBERTa}}$ | 697 | – | $61.0_{1.03}$ | $57.47_{0.56}$ |
| FinTextSen | 2486 | LLaMA $2_{0-\text{shot}}$ | 351 | 7.76 | 54.98 | 37.10 |
| | | LLaMA $2_{3-\text{shot}}$ | 90 | 2.49 | 68.89 | 3734 |
| | | $\text{FinMA}_{0-\text{shot}}$ | 65 | 1.89 | 72.31 | 44.32 |
| | | $\text{FinMA}_{3-\text{shot}}$ | 1688 | 46.16 | 68.01 | 54.42 |
| | | $\text{ChatGPT}_{0-\text{shot}}$ | 2480 | 30.57 | 30.65 | 32.17 |
| | | $\text{ChatGPT}_{3-\text{shot}}$ | 2485 | 40.27 | 40.28 | 38.6 |
| | | $\text{SuCroMoCo}_{\text{BERT}}$ | 2486 | – | $84.70_{1.46}$ | $65.02_{3.87}$ |
| | | $\text{SuCroMoCo}_{\text{RoBERTa}}$ | 2486 | – | $87.75_{1.21}$ | $66.50_{5.47}$ |
| FPB-50% | 2797 | LLaMA $2_{0-\text{shot}}$ | 421 | 5.04 | 33.49 | 31.75 |
| | | LLaMA $2_{3-\text{shot}}$ | 142 | 1.54 | 30.28 | 24.04 |
| | | $\text{FinMA}_{0-\text{shot}}$ | 1627 | 53.74 | 92.38 | 93.12 |
| | | $\text{FinMA}_{3-\text{shot}}$ | 2061 | 68.97 | 93.6 | 93.59 |
| | | $\text{ChatGPT}_{0-\text{shot}}$ | 2797 | – | 68.50 | 69.44 |
| | | $\text{ChatGPT}_{3-\text{shot}}$ | 2797 | – | 66.85 | 69.9 |
| | | $\text{SuCroMoCo}_{\text{BERT}}$ | 2797 | – | $79.40_{2.32}$ | $78.43_{2.78}$ |
| | | $\text{SuCroMoCo}_{\text{RoBERTa}}$ | 2797 | – | $80.84_{1.03}$ | $80.34_{1.12}$ |

LLaMA 2 and FinMA fall short of achieving complete coverage for valid predictions across all three datasets. Instances of numbers and blanks were observed in the invalid responses, indicating a limitation in the LLMs' ability to comprehend nuanced contextual information within financial text. Admittedly, this limitation might be attributed to the inadequacy of the provided instruction. As previously discussed, sentiment expression within financial contexts can deviate from conventional sentiment, highlighting the need for careful and context-aware instruction design. Since prompt engineering is not the primary focus of this study, we anticipate delving deeper into this issue in future research.

Among the valid classified samples, both ChatGPT and FinMA outperform $\text{SuCroMoCo}_{\text{BERT}}$ on StockSen, TweetFinSent, and FPB-50% datasets, but slightly lag behind $\text{SuCroMoCo}_{\text{RoBERTa}}$ on the StockSen and TweetFinSent datasets in terms of the macro-F1 metric. LLaMA 2 exhibits varying degrees of success in providing valid classifications and demonstrates competitive performance on the FinTextSen dataset. For the FPB-50% dataset, FinMA achieves the highest accuracy and macro-F1 scores, indicating that domain-specific knowledge is advantageous when prompting LLMs for downstream tasks.

In the 3-shot prompting scenario, LLaMA 2 experiences a decrease in both valid prediction counts and performance scores across datasets. Interestingly, both ChatGPT and FinMA also experience a performance decrease on the StockSen and TweetFinSent datasets in the 3-shot prompting scenario. This implies that, with the same instructions, the additional case demonstrations do not significantly contribute to these models' comprehension of social media financial texts. This further emphasizes the importance of meticulous instruction design to effectively guide LLMs in understanding and classifying financial sentiment. Another issue that could be further investigated to improve performance is the strategy for selecting optimal examples to insert in the prompt. In this sense, LLMs might also need more "prototypical" examples for the target labels.

On a contrasting note, our SuCroMoCo approach undertakes sentiment learning for financial texts by aligning them with prototypical texts that share the same sentiment categories in the representation space. This unique approach empowers SuCroMoCo to enhance classification performance, even when it might not fully comprehend the intricate details of the financial text content.

**Table 5**
Experimental results of the impact of prototypical examples.

(a) Classification results obtained by fine-tuning BERT on various FPB subsets with different agreement levels, while ensuring removal of duplicate instances:

| Subset | Size | ACC(%) | Macro-F1(%) |
|---|---|---|---|
| FPB-100% | 2264 | $96.85_{0.64}$ | $95.94_{0.82}$ |
| FPB-75% | 1404 | $88.97_{2.21}$ | $86.98_{2.99}$ |
| FPB-66% | 2168 | $82.07_{1.45}$ | $80.53_{1.32}$ |
| FPB-50% | 2797 | $77.39_{0.72}$ | $76.70_{0.84}$ |

(b) Classification results of $\text{SuCroMoCo}_{\text{BERT}}$ on the TweetFinSent dataset, contrasting against instances from FPB datasets with varying agreement levels:

| | Subset | ACC(%) | Macro-F1(%) |
|---|---|---|---|
| TweetFenSint | FPB-100% | $59.25_{0.39}$ | $54.54_{0.79}$ |
| | FPB-75% | $58.05_{1.00}$ | $53.59_{0.95}$ |
| | FPB-66% | $57.71_{0.93}$ | $53.33_{1.08}$ |
| | FPB-50% | $57.36_{0.52}$ | $53.41_{1.16}$ |

### 5.3. Ablation study

#### 5.3.1. Prototypical examples

We investigate the impact of prototypical examples by contrasting financial samples against different degrees of distinguishable instances and then evaluating the final classification performance of the financial samples.

The FPB datasets, as previously mentioned, are categorized based on the level of agreement among expert annotators, ranging from FPB-100% to FPB-50%. A higher degree of agreement signifies more distinguishable instances. The classification performance achieved through

BERT fine-tuning on these diverse agreement-level FPB datasets is summarized in Table 5(a). It is important to note that certain samples from the FPB-100% dataset also appear in other sets. To ensure a fair evaluation, we exclude these duplicate samples from consideration.

The results indicate that as the agreement degree decreases from 100% to 50%, there is a concurrent decline in both accuracy and macro-F1 scores achieved by BERT fine-tuning. This suggests that instances with a higher degree of agreement are easier for BERT to classify. From a representation learning perspective, this outcome also implies that BERT can generate more discriminative representations for samples with a higher degree of agreement. A similar trend is observed when employing these instances as prototypes to assist SuCroMoCo$_{\text{BERT}}$ in classifying samples from the TweetFinSent dataset, as illustrated in Table 5(b).

This finding highlights the critical role played by the prototypical examples in the classification performance of the proposed SuCroMoCo framework. Since SuCroMoCo aligns financial texts with prototypical examples in the representation space, a set of discriminative prototypical examples effectively guides the clustering of financial texts, resulting in improved classification.

### 5.3.2. Cross-momentum contrast

The cross-momentum contrast is a pivotal component of our SuCroMoCo, as it facilitates the simultaneous contrast of a financial input against multiple prototypical examples and enables bidirectional contrastive learning across financial and prototypical texts. We explore alternative approaches to contrast financial inputs against prototypical examples. Specifically, we will compare SuCroMoCo with the following approaches:

1. SupCon$_{\text{BackTrans}}$: In this approach, financial texts and prototypical examples are mixed, and the contrast is conducted between an input and its back-translation augmentation.
2. SupCon$_{\text{Prototype}}$: Here, the contrast is applied between a financial input and a single randomly selected prototypical instance from the same sentiment category.
3. SuCroMoCo$_{\text{F2P}}$ and SuCroMoCo$_{\text{P2F}}$: In these approaches, the cross contrast is solely applied in one direction, either from financial to prototypical (F2P) or from prototypical to financial (P2F).

The results presented in Table 6 provide valuable insights into the comparison of various approaches. Notably, SupCon$_{\text{Prototype}}$ significantly underperforms compared to other methods, highlighting that aligning a financial representation with just a single prototypical representation may not be adequate for capturing sentiment information effectively.

Conversely, SupCon$_{\text{BackTrans}}$ achieves superior performance to SupCon$_{\text{Prototype}}$. This improvement can be attributed to two key factors. Firstly, the mixture of financial and prototypical examples likely contributes to the performance enhancement. Secondly, the self-augmentation technique enabled by back translation serves as an additional contributing factor for the observed improvement.

When contrasting financial text solely against prototypical examples, our SuCroMoCo$_{\text{F2P}}$ achieves similar performance to SupCon$_{\text{BackTrans}}$. This seems to suggest that SuCroMoCo$_{\text{F2P}}$ can learn robust and discriminative representations, similar to what SupCon$_{\text{BackTrans}}$ accomplishes, but without the need for introducing additional augmentation samples. However, it is worth noting that when applying a reverse direction contrast (SuCroMoCo$_{\text{P2F}}$), there is a slight decrease in performance. This indicates that the choice of the contrastive anchor plays a role in the final classification performance. Conducting bidirectional cross-contrast, as done in SuCroMoCo, appears to contribute to overall performance improvement.

**Table 6**
Ablation study for the momentum contrast and contrast direction on TweetFinSent dataset.

| Model | ACC(%) | Macro-F1(%) |
|---|---|---|
| SupCon$_{\text{BackTrans}}$ | $58.64_{1.28}$ | $53.29_{0.84}$ |
| SupCon$_{\text{Prototype}}$ | $56.53_{1.70}$ | $48.52_{1.29}$ |
| SuCroMoCo$_{\text{F2P}}$ | $59.03_{0.95}$ | $53.35_{0.68}$ |
| SuCroMoCo$_{\text{P2F}}$ | $58.42_{0.92}$ | $52.96_{0.45}$ |
| SuCroMoCo | $59.25_{0.39}$ | $54.45_{0.79}$ |

## 5.4. Hyperparameters

We asses the sensitivity of SuCroMoCo to two key hyperparameters: temperature ($\tau$) and momentum coefficient ($m$). Our experiments are conducted individually on all three benchmark datasets. For temperature fine-tuning, we explore a range of values from 0.02 to 0.1 while keeping the momentum values fixed based on the results of a grid search. For temperature fine-tuning, we explore a range of values from 0.02 to 0.1 while keeping the momentum values fixed based on the results of a grid search. Following He et al. [24], we fine-tune the momentum using relatively large values such as 0.99 and 0.999. Additionally, we test the special momentum values of 0 and 1. A momentum value of 0 indicates that the parameters of the key encoder are entirely replaced by those of the trained query encoder, while a value of 1 keeps the parameters of the key encoder unchanged. The fine-tuning results for temperature are depicted in Fig. 4, while the results for momentum fine-tuning are presented in Fig. 5.

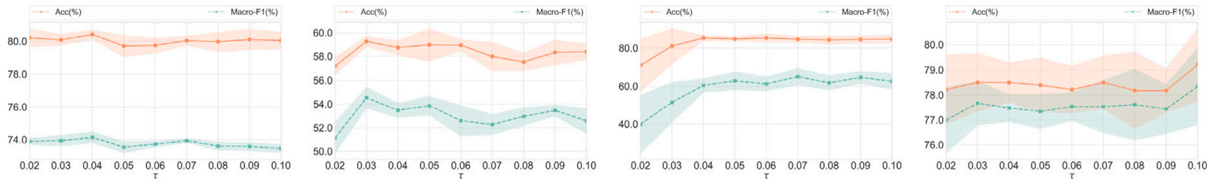
The performance analysis of SuCroMoCo in response to variations in the temperature parameter ($\tau$) is presented in Fig. 4. Notable insights are observed across different datasets, revealing the impact of temperature on SuCroMoCo's performance.

On the StockSen dataset, subtle fluctuations are observed, with the mean accuracy hovering around 80.0% and the mean macro-F1 score approximately at 74.0%. The optimal configuration is achieved at a temperature of 0.04, yielding the highest mean accuracy score of 80.42% and a macro-F1 score of 74.15%. Conversely, the FinTextSen dataset exhibits a clear rise in both accuracy and macro-F1 scores when $\tau$ is lower than 0.04. Beyond this point, the accuracy score stabilizes around 80%, while the macro-F1 score slightly fluctuates between 66.50% and 60.0%.

The TweetFinSent dataset displays more pronounced fluctuations, with mean accuracy scores ranging from around 57.0% to about 59.0%, and macro-F1 scores varying from approximately 51.0% to 55.0%. The temperature value of 0.03 leads to the best performance, achieving an average accuracy of 59.25% and a macro-F1 score of 54.54%. For the FPB dataset, mean accuracy and macro-F1 scores remain relatively stable at around 78.5% and 77.5%, respectively, until $\tau$ drops below 0.09. A notable performance boost occurs when $\tau = 0.1$, resulting in mean accuracy and macro-F1 scores reaching their highest values at 79.40% and 78.43%, respectively.

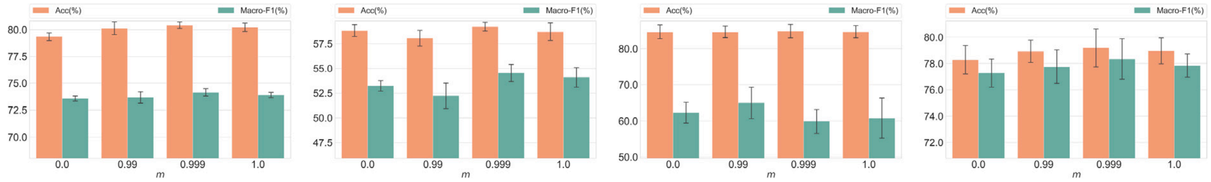
These findings restate the importance of selecting an appropriate temperature parameter for SuCroMoCo, emphasizing its sensitivity to dataset characteristics. Optimal performance is achieved when tailoring the temperature parameter to suit the specific nuances of each dataset.

Fig. 5 illustrates the fine-tuning results of the momentum coefficient ($m$) on the four datasets, providing insights into how SuCroMoCo's performance responds to variations in this parameter. If we compare these results to the fine-tuning of the temperature parameter, a notable observation is that SuCroMoCo's performance remains relatively stable across the four datasets as $m$ changes. For the FinTextSent dataset, the accuracy scores remain constant, while the macro-F1 score slightly reaches its highest value at $m = 0.99$. This stability in performance suggests that SuCroMoCo is less sensitive to variations in $m$ for this specific dataset. On the StockSen, TweetFinSent, and FPB-50% datasets, the optimal performance for SuCroMoCo is consistently achieved at



(a) StockSen ($m = 0.999$) (b) TweetFinSent ($m = 0.999$) (c) FinTextSen ($m = 0.99$) (d) FPB-50% ($m = 0.999$)

**Fig. 4.** The impact of different temperature $\tau$ values on classification performance on StockSen, TweetFinSent, FinTextSen, and FPB-50% datasets, respectively. For each dataset, the momentum $m$ value is set to the optimal value and fixed.



(a) StockSen ($\tau = 0.04$) (b) TweetFinSent ($\tau = 0.03$) (c) FinTextSen ($\tau = 0.07$) (d) FPB-50% ($\tau = 0.1$)

**Fig. 5.** The impact of different momentum $m$ values on classification performance on StockSen, TweetFinSent, FinTextSen, and FPB-50% datasets, respectively. For each dataset, the temperature $\tau$ value is set to the optimal value and fixed.

$m = 0.999$. This indicates that setting $m$ to a high value is favorable for aligning query and key representations in these datasets.

It is interesting to note that SuCroMoCo still yields meaningful outcomes when $m$ is set to 0, in contrast to the findings by He et al. [24], who reported that the experiments failed under this condition. The bidirectional cross-momentum contrast mechanism in SuCroMoCo likely contributes to this difference. Nevertheless, SuCroMoCo performs slightly worse when $m$ is set to 0 compared to when $m$ is set to 1, suggesting that maintaining the key encoder unchanged does not provide a significant benefit for aligning query and key representations. A small $m$ value is preferable to facilitate the rapid change of the key encoder, thus emphasizing the nuanced impact of the momentum coefficient on SuCroMoCo's performance.

## 6. Conclusion

In this paper, we presented SuCroMoCo, a novel framework for financial sentiment analysis that combines and leverages supervised contrastive learning and cross-momentum contrast. One of the main strengths of SuCroMoCo is its capability to align financial text representations with prototypical representations based on the sentiment categories. This alignment proves highly effective, even in cases where the pre-trained language models have limited comprehension of the content of financial texts. As a result, SuCroMoCo significantly improves classification performance. Through extensive experiments, we demonstrate that SuCroMoCo outperforms various existing approaches in the field of financial sentiment analysis.

In addition, we thoroughly investigated the impact of various factors on SuCroMoCo's performance, including the discriminative level of prototypical examples, the choice of hyperparameters such as temperature and momentum, and the direction of cross-momentum contrast. The findings from these experiments not only provided valuable insights about the internal mechanisms of SuCroMoCo but also offered practical guidance for implementation.

Nevertheless, it is important to acknowledge the limitations of SuCroMoCo. The requirement for prototypical examples to share the same label space as financial samples restricts its applicability to datasets in which label mapping is feasible. This means that SuCroMoCo may provide an effective solution for tasks such as financial sentiment analysis, where the label space is somehow conventional (two or three classes, typically corresponding to the classical sentiment polarities); however its application to other types of tasks and datasets may be more challenging.

In conclusion, SuCroMoCo provides a straightforward and efficient solution for improving the performance of financial sentiment analysis, even without complete comprehension of financial text content. Future research can focus on expanding the framework's applicability and assessing its effectiveness in various financial tasks and contexts.

## CRediT authorship contribution statement

**Bo Peng:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Emmanuele Chersoni:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis, Conceptualization. **Yu-yin Hsu:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition. **Le Qiu:** Writing – review & editing, Validation, Data curation. **Chu-ren Huang:** Writing – review & editing, Supervision, Resources, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The code and datasets can be found in https://github.com/PengBO-O/SuCroMoCo.

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

## Acknowledgments

The authors acknowledge the support from the project "Analyzing the semantics of Transformers representations for financial natural language processing" (ZVYU), funded by the Faculty of Humanities of the Hong Kong Polytechnic University (PolyU-UGC).

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2024.111683.

## References

[1] S.W. Chan, M.W. Chong, Sentiment analysis in financial texts, Decis. Support Syst. 94 (2017) 53–64, http://dx.doi.org/10.1016/j.dss.2016.10.006.

[2] X. Li, X. Huang, X. Deng, S. Zhu, Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information, Neurocomputing 142 (2014) 228–238, http://dx.doi.org/10.1016/j.neucom.2014.04.043.

[3] R. Gupta, M. Chen, Sentiment analysis for stock price prediction, in: 2020 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR, IEEE, Shenzhen, Guangdong, China, 2020, pp. 213–218, http://dx.doi.org/10.1109/MIPR49039.2020.00051.

[4] M. Cecchini, H. Aytug, G.J. Koehler, P. Pathak, Making words work: Using financial text as a predictor of financial events, Decis. Support Syst. 50 (1) (2010) 164–175, http://dx.doi.org/10.1016/j.dss.2010.07.012.

[5] P. Craja, A. Kim, S. Lessmann, Deep learning for detecting financial statement fraud, Decis. Support Syst. 139 (2020) 113421, http://dx.doi.org/10.1016/j.dss.2020.113421.

[6] P.C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, J. Finance 62 (3) (2007) 1139–1168, http://dx.doi.org/10.1111/j.1540-6261.2007.01232.x.

[7] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: The AZFin text system, ACM Trans. Inf. Syst. 27 (2) (2009) 1–19, http://dx.doi.org/10.1145/1462198.1462204.

[8] C. Nopp, A. Hanbury, Detecting risks in the banking system by sentiment analysis, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 591–600, http://dx.doi.org/10.18653/v1/D15-1071.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, 2019, http://dx.doi.org/10.48550/arXiv.1907.11692, arXiv:1907.11692.

[11] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training 12.

[12] K. Clark, M.-T. Luong, Q. Le, C.D. Manning, Pre-training transformers as energy-based cloze models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 285–294, http://dx.doi.org/10.18653/v1/2020.emnlp-main.20.

[13] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification? in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18, Springer, 2019, pp. 194–206.

[14] L. Zhao, L. Li, X. Zheng, J. Zhang, A BERT based sentiment analysis and key entity detection approach for online financial texts, in: 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD, IEEE, Dalian, China, 2021, pp. 1233–1238, http://dx.doi.org/10.1109/CSCWD49262.2021.9437616.

[15] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune BERT for text classification? in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2019, pp. 194–206, http://dx.doi.org/10.1007/978-3-030-32381-3_16.

[16] K. Mishev, A. Gjorgjevikj, I. Vodenska, L.T. Chitkushev, D. Trajanov, Evaluation of sentiment analysis in finance: From lexicons to transformers, IEEE Access 8 (2020) 131662–131682, http://dx.doi.org/10.1109/ACCESS.2020.3009626.

[17] Y. Yang, M.C.S. UY, A. Huang, FinBERT: A pretrained language model for financial communications, 2020, arXiv:2006.08097.

[18] D. Araci, FinBERT: Financial sentiment analysis with pre-trained language models, 2019, arXiv:1908.10063 [cs]. arXiv:1908.10063.

[19] R.S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, D. Yang, WHEN FLUE MEETS FLANG: Benchmarks and large pre-trained language model for financial domain.

[20] B. Peng, E. Chersoni, Y.-Y. Hsu, C.-R. Huang, Is domain adaptation worth your investment? Comparing BERT and FinBERT on financial tasks, in: Proceedings of the Third Workshop on Economics and Natural Language Processing, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 37–44, http://dx.doi.org/10.18653/v1/2021.econlp-1.5.

[21] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, P. Takala, Good debt or bad debt: Detecting semantic orientations in economic texts, J. Assoc. Inf. Sci. Technol. 65 (4) (2014) 782–796, http://dx.doi.org/10.1002/asi.23062.

[22] F. Xing, L. Malandri, Y. Zhang, E. Cambria, Financial sentiment analysis: An investigation into common mistakes and silver bullets, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 978–987, http://dx.doi.org/10.18653/v1/2020.coling-main.85.

[23] Y. Pei, A. Mbakwe, A. Gupta, S. Alamir, H. Lin, X. Liu, S. Shah, TweetFinSent: A dataset of stock sentiments on Twitter, in: Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing, FinNLP, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 37–47.

[24] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Seattle, WA, USA, 2020, pp. 9726–9735, http://dx.doi.org/10.1109/CVPR42600.2020.00975.

[25] P. Khosla, Y. Tian, P. Teterwak, C. Wang, P. Isola, A. Maschinot, D. Krishnan, A. Sarna, Supervised Contrastive Learning 13.

[26] B. Gunel, J. Du, A. Conneau, V. Stoyanov, Supervised contrastive learning for pre-trained language model fine-tuning, 2021, arXiv:2011.01403.

[27] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, 2020.

[28] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910, http://dx.doi.org/10.18653/v1/2021.emnlp-main.552.

[29] T. Loughran, B. Mcdonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, J. Finance 66 (1) (2011) 35–65, http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x.

[30] C.-C. Chen, H.-H. Huang, H.-H. Chen, NTUSD-Fin: A market sentiment dictionary for financial social media data applications, in: Proceedings of the 1st Financial Narrative Processing Workshop, FNP 2018, 2018, pp. 37–43.

[31] F.Z. Xing, F. Pallucchini, E. Cambria, Cognitive-inspired domain adaptation of sentiment lexicons, Inf. Process. Manage. 56 (3) (2019) 554–564, http://dx.doi.org/10.1016/j.ipm.2018.11.002.

[32] R. Ren, D.D. Wu, T. Liu, Forecasting stock market movement direction using sentiment analysis and support vector machine, IEEE Syst. J. 13 (1) (2019) 760–770, http://dx.doi.org/10.1109/JSYST.2018.2794462.

[33] T. Pranckevičius, V. Marcinkevičius, Comparison of naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification, Baltic J. Modern Comput. 5 (2) (2017) http://dx.doi.org/10.22364/bjmc.2017.5.2.05.

[34] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735, arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

[35] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 160–167.

[36] M.S. Akhtar, A. Kumar, D. Ghosal, A. Ekbal, P. Bhattacharyya, A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 540–546, http://dx.doi.org/10.18653/v1/D17-1057.

[37] E. Shijia, L. Yang, M. Zhang, Y. Xiang, Aspect-based financial sentiment analysis with deep neural networks, in: WWW (Companion Volume), 2018, pp. 1951–1954.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[39] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines, 2021, arXiv:2006.04884.

[40] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N.A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360, http://dx.doi.org/10.18653/v1/2020.acl-main.740.

[41] D. Grangier, D. Iter, The trade-offs of domain adaptation for neural language models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3802–3813, http://dx.doi.org/10.18653/v1/2022.acl-long.264.

[42] B. Peng, E. Chersoni, Y.-Y. Hsu, C.-R. Huang, Discovering Financial Hypernyms by Prompting Masked Language Models, in: Proceedings of the LREC Workshop on Financial Narrative Processing Workshop, 2022.

[43] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[44] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2019, arXiv:1807.03748.

[45] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, W. Xu, ConSERT: A contrastive framework for self-supervised sentence representation transfer, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 5065–5075, http://dx.doi.org/10.18653/v1/2021.acl-long.393.

[46] Y.-S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljacic, S.-W. Li, S. Yih, Y. Kim, J. Glass, DiffCSE: Difference-based contrastive learning for sentence embeddings, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4207–4218, http://dx.doi.org/10.18653/v1/2022.naacl-main.311.

[47] X. Wu, C. Gao, L. Zang, J. Han, Z. Wang, S. Hu, ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3898–3907.

[48] Z. Dai, B. Cai, J. Chen, UniMoCo: Unsupervised, semi-supervised and fully-supervised visual representation learning, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics, SMC, 2022, pp. 3099–3106, http://dx.doi.org/10.1109/SMC53654.2022.9945500.

[49] Q. Chen, R. Zhang, Y. Zheng, Y. Mao, Dual contrastive learning: Text classification via label-aware data augmentation, 2022, arXiv:2201.08702.

[50] L. Wang, W. Zhao, J. Liu, Aligning cross-lingual sentence representations with dual momentum contrast, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3807–3815, http://dx.doi.org/10.18653/v1/2021.emnlp-main.309.

[51] Q. Tan, R. He, L. Bing, H.T. Ng, Domain generalization for text classification with memory-based supervised contrastive learning, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6916–6926.

[52] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742, http://dx.doi.org/10.1109/CVPR.2018.00393.

[53] R. Wang, Z. Wu, Z. Weng, J. Chen, G.-J. Qi, Y.-G. Jiang, Cross-domain contrastive learning for unsupervised domain adaptation, IEEE Trans. Multimed. 25 (2023) 1665–1673, http://dx.doi.org/10.1109/TMM.2022.3146744.

[54] K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, B. Davis, SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news, in: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEVal-2017, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 519–535, http://dx.doi.org/10.18653/v1/S17-2089, URL https://aclanthology.org/S17-2089.

[55] T. Daudert, P. Buitelaar, S. Negi, Leveraging news sentiment to improve microblog sentiment classification in the financial domain, in: Proceedings of the First Workshop on Economics and Natural Language Processing, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 49–54, http://dx.doi.org/10.18653/v1/W18-3107.

[56] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 1631–1642.

[57] F. Karl, A. Scherp, Transformers are short-text classifiers, in: A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A.M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer Nature Switzerland, Cham, 2023, pp. 103–122.

[58] H. Wu, X. Shi, Adversarial soft prompt tuning for cross-domain sentiment analysis, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2438–2447, http://dx.doi.org/10.18653/v1/2022.acl-long.174.

[59] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent Abilities of Large Language Models, 2022, arXiv preprint arXiv:2206.07682.

[60] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C.C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P.S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E.M. Smith, R. Subramanian, X.E. Tan, B. Tang, R. Taylor, A. Williams, J.X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023, arXiv:2307.09288.

[61] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, PIXIU: A large language model, instruction data and evaluation benchmark for finance, 2023, arXiv:2306.05443.

[62] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, 2023, arXiv:2302.13971.

[63] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022, URL https://openreview.net/forum?id=nZeVKeeFYf9.

[64] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, GPT3. Int8 (): 8-bit matrix multiplication for transformers at scale, in: Advances in Neural Information Processing Systems, 35, 2022, pp. 30318–30332.

[65] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.

[66] P. Micikevicius, S. Narang, J. Alben, G.F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, H. Wu, Mixed precision training, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018, URL https://openreview.net/forum?id=r1gs9JgRZ.

[67] J.R. Hermans, G. Spanakis, R. Möckel, Accumulated gradient normalization, in: Proceedings of the 9th Asian Conference on Machine Learning, PMLR, 2017, pp. 439–454.